DOCUMENT RESUME

ED 193 939 PL 011 904

AUTHOR Benzon, William L.
TITLE Computational Linguistics and Discourse Analysis.
PUB DATE 79
NOTE 13p.: Paper presented at the Annual Meeting of the North-East Modern Language Association (Hartford, CT, March 29-31, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computational Linguistics: *Discourse Analysis: *Linguistic Theory: Models: Semantics

ABSTRACT

The profound use of the computer in discourse analysis must employ a theory of discourse comprehension and production with which to conduct the analysis. Models currently employed in computational linguistics have a semantic basis and are goal-directed. The basic model is an associative cognitive network. The basic inventory of concepts of the system is given in the systemic network, which is organized into paradigmatic, syntagmatic, and componential structures. Since events happen in particular places at particular times, there is also an episodic structure. The gnomonic system defines abstract concepts over episodes. According to Phillips (1975), discourse coherence must be considered on two levels, the episodic and the gnomic. A discourse which engenders episodic and/or gnomonic expectations which are not then fulfilled is incoherent. A lower limit on coherence may be defined as a discourse so ill-formed that it makes no sense even to its creator. The upper limit on coherence is set by the most powerful creative minds. Between the two limits, discourse analysis, from the point of view of the computational linguist, probably requires nothing less than a full-blown computational theory of the human mind. (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193939

COMPUTATIONAL LINGUISTICS AND DISCOURSE ANALYSIS

William L. Benzon

Language, Literature and Communications
Rensselaer Polytechnic Institute
Troy, New York 12181

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. L. Benzon

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Loosely speaking computational linguistics is any use of the computer in connection with the study of natural languages. But there is a superficial use of the computer and a profound use. The superficial use deals only with the surface of language and produces concordances, word counts, and statistical analyses. The profound use treats the computer as a device for simulating theories about the production and comprehension of discourse. This paper is concerned only with the profound use (for a review of the field see Benzon and Hays 1976).

I have only one point to make: The profound analysis of discourse must employ a theory of discourse comprehension and production with which to conduct the analysis. There is no inductive analytical procedure which one can apply to texts and somehow magically "come up with" a theory of discourse. Rather, one must first "come up with" a theory of discourse, no matter how crude the theory might be (and our present theories are very crude), and then see how well it performs with a body of texts. To evaluate the model's performance one can simulate the model on a computer and one can use it as a basis for psycholinguistic experimentation (Norman et al. 1975, Kintsch 1974, Thorndyke 1977). One then creates a better model and it must, in turn, be evaluated.

FL 011 904

The models currently employed in computational linguistics have two characteristics which are important:

1) They start with a semantic basis. Discourse is produced by operation on a semantic base and it is comprehended by assimilation into that semantic base.
2) Time is intrinsic to the model. Discourse production and comprehension and production is goal-directed, making constant use of projections and anticipations of upcoming elements in the speech stream.

To analyse a body of texts one must create a model of the semantic base underlying the texts and of the operations on that base which generate discourse. The analysis consists in the application of the model to the body of texts.

Within this paradigm a theory of discourse is really a theory of the inner structure and processes of the computational model. One is concerned, not with texts per se, but with the processes by which people create and understand texts. The model which I describe here has been developed by David G. Hays and his students at SUNY Buffalo.

COGNITIVE NETWORKS

The basic model is an associative cognitive network. Imagine a spider's web. The junctions between threads are called nodes while the threads are arcs or links. Each node is a concept while the arcs specify the relationships which exists between the concepts at either ends of the arcs. Discourse is produced by generating a path (or paths) through the network (imagine planning a trip using a roadmap) and it is comprehended by assimilating a particular path into the network.

Both semantics and syntax are embedded within a network structure. (In this model pragmatics is essentially higher order semantics, see Bloom and Hays, in preparation.) The syntactic network operates on the semantic network. That is, processes in the semantic network are controlled by the syntactic network.

Semantics is relational and, in a sense, spatial. The meaning of a given node is specified by the place which it occupies in the entire network. That place is given by the arcs which impinge on the node. Syntax is temporal, placing one item after another in the speech stream. The job of the syntactic network is to mediate between the spatial relationality of the semantic network and the linear unfolding of the speech chain in which only one term of a complex relational nexus can be given at a time.

THE SYSTEMIC NETWORK

The basic inventory of concepts of the system is given in the systemic network, which is organized into paradigmatic, syntagmatic, and componential structure.

There are two types of paradigms: substantive and functional. Items are organized into substantive paradigms according to their sensory attributes. One such paradigm has plant at its root. Tree, grass, herb, vine, and bush are all varieties of plant and oak, pine, maple, sycamore, palm, ginko, etc. are, in turn, varieties of tree. (Such paradigms have been examined on a cross-cultural basis, see Berlin, Breedlove and Raven, 1973.)

Functional paradigms organize items according to their use. Foods are those plants and animals which can be eaten. And foods can be classified according to their methods of preparation, the ways in which they are eaten, or their place in the menue.

Syntagmatic structure gives the relationship between properties, entities, events, and plans. Redness, roundness, smoothness, are properties which participate in the entity apple. When it falls from the tree the apple is participating in the event fall. And when it is thrown at Johnny's head the moving apple is participating in a plan, hit Johnny on the head.

Componential structure relates parts to wholes. A tree consists of trunk, roots, branches, and foliage. Your typical bird has a head, a neck, a body, two wings, two legs and a tail. The act of hitting a baseball includes watching the ball, swinging the bat to the ball, the follow-through, and watching the ball sail over the fence.

The relationship between any two concepts in the systemic network can be established through tracing the path between the two nodes. Imagine the network as a fisherman's net. The cords are the arcs and the knots are the concepts. Grab the two nodes in question and pull tight; you have found the shortest path between the nodes. The path between tree and applesauce would consist of three links: 1) a paradigmatic link between tree and apple tree, 2) a component link between apple tree and apple, and 3) a syntagmatic link between apple and mash (the process by which applesauce is created). A paradigmatic link in a functional paradigm would link applesauce to food. All of these links (and some others) would be traversed in producing or comprehending the sentence: Applesauce is a food created by mashing the fruit of the apple tree.

EPISODIC STRUCTURE

Events happen in particular places at particular times. For this we have episodic structure. It is all well and good to talk of apple mashing, but what of that particularly fine apple mashing the Walton's had to celebrate

the publication of John Boy's first short story? That happened at a particular time and in a particular place and so that record is kept in the episodic store. The episodic store is thus the system's historical archive.

Typical episodic structures form the basis of frames (Minsky 1975) or scripts and plans (Schank and Abelson 1977). The creation of good applesauce is actually a moderately complicated affair, involving the coordination of events in several different places at several different times. First one must get the apples (from the orchard or from the grocer); that happens in one place. Then the apples must be moved to another place where they are washed. They are then moved (but perhaps not so far as the first time) to a place where they are peeled (an optional step) and cored (not so optional). After this they are simmered (a slightly different place) and then mashed (yet another place). And that, roughly, is how you make applesauce. It is too complicated to be handled by basic systemic structure. Rather, it is a spatio-temporal organization of systemic structures.

The same episodic frame which is used to perform some activity can also be used to produce and comprehend discourse about that activity. I use my applesauce-making frame to create my little story about the applesaucing of John Boy and you use your applesauce-making frame to comprehend my story. Even as John Boy is in the orchard picking the apples you are using the frame to anticipate the next step in the story, and then the step after that. You match my story against your internalized applesauce-making frame. But when Grandpa empties a quart of vodka into the saucepan your attention is aroused - that certainly is not in the applesauce-frame. And so you must now embed your consequences-of-drink frame in the applesauce frame. This causes you to anticipate that, at some point in the story, John Boy is going to get drunk and do something he might regret.

Well, not quite. John Boy does get roaring drunk. And he reveals that he had plagerized the story from a friend of his. He's been feeling guilty for weeks, but now that he's told the truth he feels better and he's going to tell his friend what he did and make sure that the publisher lists the name of the real author. He asks all to forgive him for what he's done.

Thus, "The Apple-Bousing of John Boy" is a story with a moral: only in sincere repentance do the guilty find relief. This story involves abstract concepts: guilt, repentance, justice. The agent of injustic feels guilt and can find relief only in repentance.

To understand this we have to consider the next level of the system.

THE GNOMONIC SYSTEM

The gnomonic system defines abstract concepts over episodes. The story of John Boy is a particular example of repentance. There are many other such stories. Within this particular system all abstract concepts are defined over sets of episodes containing exemplary stories (Benzon 1976, 1978, ; Hays 1973, 1976; Phillips 1975, in press; White 1975). It is particularly important to note that stories which themselves define a certain abstract concept can contain abstract concepts. Thus one can talk of a first rank abstraction as one defined over stories containing no abstract concepts. A second rank abstraction is defined over stories which contain at least one first rank abstration. By continuing this process, which is recursive, it is possible to build up concepts of indefinitely high abstractive rank. There is some reason to believe that cultural evolution proceeds in just this way (Benzon 1978).

Let us consider another example.

3) Mary went into the woods and saw some pretty mushrooms. She picked them and returned home where she ate them. Shortly thereafter she

became violently ill. Finally, she died.

4) Billy was playing in the yard. A big hairy spider came up and bit him. Not too long afterward he became violently ill and he was unconscious for three days before he finally revived.

Both of these stories involve poison. But we have two senses of poison. The physical substance, the mushroom, the spider's venom, is poison by functional definition. Just as something is food by virtue of its capacity to be eaten, so something is a poison by virtue of its capacity to fill a certain role in stories such as 3) and 4) above.

But we also have an abstract concept of poison which emerges only through consideration of the whole pattern. Abstractly considered,

5) Poison is an evil spirit which causes a person's soul to leave his body, temporarily or permanently.

Abstract poison is an ineffable substance which exists in certain physical substances (namely, those functionally defined as poisons) which causes them to have certain effects. Statement 5) is a rationalization of abstract poison, it is an attempt to explain how poisons (functionally defined) have their effect. Other elements of that rationalization must also be abstractions (soul, and evil spirit (poison is just a variety of evil spirit)).

Thus, associated with every abstract concept we have the set of exemplary episodes which illustrate the concept and which provides the primary definitional basis for the concept, and the rationalization, which attempts to explicate the concept and which is, as such, the secondary definitional basis of the concept. Notice that this account is consonant with Thomas Kuhn's notion of a paradigm (1970). The primary definitional basis of a Kuhnian paradigm consists of exemplary experiments and problems. The

explicit rules of science are secondary to those examples. Those explicit rules are, in my terminology, rationalizations.

DISCOURSE COHERENCE

According to Brian Phillips (1975, in press) discourse coherence must be considered on two levels, the episodic and the gnomonic. At the episodic level temporal, causal, and spatial relationships must form a coherent pattern. One can't have John Boy picking apples in Twentieth Century Europe and mashing them in Nineteenth Century Africa - at least not in the humble sort of story I described. At the gnomonic level a discourse must have a theme, that is, it must be an instance of some abstract concept. A discourse can be coherent at the episodic level without having any significant gnomonic structure - a straight historical chronical (and I do mean very straight) would be such a discourse. And gnomonic patterning may absorb apparent anomalies at the episodic level - a rather staid science fiction story can use time travel to have John Boy mash the apples even before he's been born and a writer of contemporary metafiction might use a similar anomaly for a different effect.

A discourse which engenders episodic and/or gnomonic expectations which are not then fulfilled is incoherent. However, it is rarely the case that all the information needed to understand a discourse is present explicitly in the text. Much must be inferred. Consequently it is possible that a discourse which is coherent for one person is incoherent for another. If one doesn't have the knowledge necessary to make the proper inferences on the basis of the information presented in a discourse, then the discourse will appear to be incoherent without in fact being so. Coherence is a property of the relationship between a given discourse and the semantic

base into which that discourse is being assimilated.

This is a fairly relativistic notion of coherence, but it isn't quite equivalent to asserting that any discourse is coherent to someone. Consider the fairly frequent situation where someone will make some notes, write a few paragraphs or so, and then come back to that discourse a few hours, days, weeks, etc. later and find the discourse completely unintelligible. Here is a case of a discourse being incoherent in relationship to the semantic base from which it was produced. For that matter, much of the difficulty of writing coherent prose is precisely in the process of making discourse coherent to the author (i.e. to the semantic base from which the discourse is generated). Thus we are not left with the uninteresting notion that any discourse is coherent to someone. Some discourses are so ill-formed that they make sense to no one, not even their creators.

If that defines a lower limit to discourse coherence, then perhaps we might consider what an upper limit might be like. It is no secret that literary critics have widely divergent views on the meaning and significance of literary texts. Norman Holland explains this by the concept of identity theme (Holland 1975). Different people have different identity themes (i.e. personalities) and so read texts differently; each reads according to his own identity theme. Presumably differences in identity theme could be translated into differences in semantic bases, so I have no quarrel with Holland. But I want to make a different suggestion.

Most of the texts studied by professional students of literature were written by people whose mental and creative powers are probably greater than those of their professional students. Thus what was coherent to the artist might be incoherent to the critic whose powers vis-à-vis the text are like those of the blind men vis-à-vis the elephant. This situation

would also lead to critical chaos. We accept this sort of principle when we assign grade levels to texts intended for school children. Why not apply it to ourselves? Perhaps we are 21st graders reading 25th grade texts.

The upper limits to discourse coherence are thus set by the most powerful creative minds. Between the upper and the lower limits we have a cultural community, a group of individuals whose various discourses are mutually coherent in varying degrees. A discourse which falls below the lower limit is coherent to no one. But the upper limit defines the degree to which apparently conflicting discourses can become mutually coherent through higher level patterning.

CONCLUSION

It is probably the case that, for the computational linguist, discourse analysis requires nothing less than a full-blown computational theory of the human mind. That is a tall order. And we are not close to filling it. Indeed, if the human mind does in fact posess the recursive abstraction power this model attributes to it, then the mind will always outstrip our efforts to model it (for it will be constructing the model). But there is much to be learned in attempting to create a computational theory of the human mind. And the tools with which to create that theory are available to those who would use them. The field of computational linguistics is immature and rich in promise.

References

Benzon, William L. Cognitive Networks and Literary Semantics. MLN 91: 952 - 982, 1976.

Benzon, William L. Cognitive Science and Literary Theory. Unpublished Doctoral Dissertation, SUNY at Buffalo, 1978.

Benzon, William L. and David G. Hays. Computational Linguistics and the Humanist. Computers and the Humanities 10: 265 - 274, 1976.

Berlin, Brent, Dennis E. Breeklove, and Peter H. Raven. General Principles of Classification and Nomenclature in Folk Biology. American Anthropologist 75: 214 - 242, 1973.

Bloom, David, and David G. Hays. Designation in English. In press.

Hays, David G. The Meaning of a Term is a Function of the Theory in Which It Occurs. SIGLASH Newsletter 6, No. 4: 8 - 11, 1973.

Hays, David G. On "Alienation": An Essay in the Psycholinguistics of Science. In R. Felix Geyer abd David R. Schweitzer, eds. Theories of Alienation. Martinus Nijhoff, 1976, 169 - 187.

Holland, Norman. Five Readers Reading. New Haven: Yale University Press, 1975.

Kintsch, Walter. The Representation of Meaning in Memory. Hillsdale: Lawrence Erlbaum, 1974.

Kuhn, Thomas. The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1970.

Minsky, Marvin. A Framework for Representing Knowledge. In P.H. Winston, ed. The Psychology of Computer Vision. New York: McGraw-Hill, 1975.

Norman, Donald A., and David E. Rumelhard and the LNR Research Group. Explorations in Cognition. San Francisco: Freeman, 1975.

Phillips, Brian. Topic Analysis. Unpublished Doctoral Dissertation, SUNY Buffalo, 1975.

Phillips, Brian. A Model for Knowledge and Its Application to Discourse Analysis. American Journal of Computational Linguistics, in press.

Schank, Roger, and R. Abelson. Scripts Plans Goals and Understanding. Hillsdale: Lawrence Erlbaum, 1977.

Thorndyke, Perry. Pattern-Directed Processing of Knowledge from Texts. Rand Paper P-5806, May 1977.

White, Mary J. Cognitive Networks and Worldview: The Metaphysical Terminology of a Millenarian Community. Unpublished Doctoral Disseration, SUNY Buffalo, 1975.